A

Express Mail transmittal

01/27/99 Jc530 U.S. PTO

UTILITY PATENT APPLICATION TRANSMITTAL

(only for new non-provisional applications under 37 CFR 1.53(b))

Attorney Docket No.	**Beutnagel 3-12-9**
First Named Inventor or Application Identifier	**Mark Beutnagel**
Title	**Advanced TTS for Facial Animation**
Express Mail Label no.	**EM164542418US**

JC523 U.S. PTO 09/238224 01/27/99

To: **Assistant Commissioner for Patents**
Box Patent Application
Washington D.C. 20231

APPLICATION ELEMENTS	ACCOMPANYING APPLICATION PARTS
☒ Fee Transmittal Form (original and duplicate)	☐ Assignment
☒ Specification Total Pages **13** title cross reference to related applications (e.g. provisional application) background summary brief description of the drawings (if filed) detailed description claims abstract	☐ Recordation form ☒ Power of Attorney ☒ Postcard ☐ Small entity statement ☐ Certified copy of priority documents ☐ Information disclosure statement ☐ Copies of IDS citations ☐ 37 CFR 3.73(b) Statement ☒ check ☐ Other
☒ Drawing(s) Total Pages **2**	
☒ Declaration Total Pages **3**	
a. ☐ Newly executed	
b. ☐ Copy from a prior application (37 CFR 1.63(d)) (for continuations/divisionals with section below filled out)	
☐ Deletion of Inventor(s) Signed Statement attached deleting inventor(s) named in the prior application. 37 CFR 163 (d)(2) and 1.33(b).	
☐ Incorporation by reference (usable if Declaration is a copy): The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied, is considered as being part of the disclosure of the accompanying application is hereby incorporated by reference herein.	
☐ Other	

If a CONTINUING APPLICATION, check appropriate box and supply the requisite information:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior Application No:

CORRESPONDENCE ADDRESS

☐ Customer Number or Bar Code Label	(insert Customer No. or Attach bar code label here)	☒ Correspondence Address below

NAME	**Samuel H. Dworetsky**	
ADDRESS	**AT&T Corp. P.O. Box 636** **Middletown, NJ 07748-4801**	
COUNTRY	**United States**	FAX **(732) 957-5505**

SIGNATURE OF APPLICANT ATTORNEY, OR AGENT

Name	**Henry T. Brendzel**	Reg. No. **26,844**
Telephone	**(973) 467-2025**	
Signature	Henry Brendzel	Date 1/27/99

I hereby certify that this Application is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington D.C. 20231.

1/27/99	**Henry Brendzel**	Henry Brendzel
Date of Deposit	(Printed Name of Person Mailing Paper)	(Signature of Person Mailing Paper)

Advanced TTS For Facial Animantion

Reference to a Related Application

This invention claims the benefit of provisional application No. 60/073185, filed January 30, 1998, titled "Advanced TTS For Facial Animation," which is incorporated by reference herein, and of provisional application No. 60/082,393, filed April 20, 1998, titled "FAP Definition Syntax for TTS Input." This invention is also related to a copending application, filed on even date hereof, titled "FAP Definition Syntax for TTS Input," which claims priority based on the same provisional applications.

Background of the Invention

The success of the MPEG-1 and MPEG-2 coding standards was driven by the fact that they allow digital audiovisual services with high quality and compression efficiency. However, the scope of these two standards is restricted to the ability of representing audiovisual information similar to analog systems where the video is limited to a sequence of rectangular frames. MPEG-4 (ISO/IEC JTC1/SC29/WG11) is the first international standard designed for true multimedia communication, and its goal is to provide a new kind of standardization that will support the evolution of information technology.

When synthesizing speech from text, MPEG 4 contemplates sending a stream containing text, prosody and bookmarks that are embedded in the text. The bookmarks provide parameters for synthesizing speech and for synthesizing facial animation. Prosody information includes pitch information, energy information, etc. The use of FAPs embedded in the text stream is described in the aforementioned copending application, which is incorporated by reference. The synthesizer employs the text to develop phonemes and prosody information that are necessary for creating sounds that corresponds to the text.

The following illustrates a stream that may be applied to a synthesizer, following the application of configuration signals. FIG. 1 provides a visual representation of this stream.

Syntax:	**# of bits**
TTS_Sentence() {	
TTS_Sentence_Start_Code	32
TTS_Sentence_ID	10

```
Silence                                                    1
if (Silence)
    Silence_Duration                                       12
else    {
    if (Gender_Enable)
        Gender                                             1
    if (Age_Enable)
        Age                                                3
    if (!Video_Enable & Speech_Rate_enable)
        Speech_Rate                                        4
    Length_of_Text                                         12
    For (j=0; j<=Length_of_Text; j++)
        TTS_Text                                           8
    if (Video_Enable) {
        if (Dur_Enable) {
            Sentence_Duration                              16
            Postion_in_Sentence                            16
            Offset                                         10
            }
        }
    if (Lip_Shape_Enable) {
        Number_of_Lip_Shape                                10
        for (j=0; j<Number_of_Lip_Shape; j++) {
            If (Prosody_Enable) {
                If (Dur_Enable)
                    Lip_Shape_Time_in_Sentence             16
                Else
                    Lip_Shape_Phoneme_Number_in_Sentence   13
            }
            else
                Lip-Shape_Letter_Number_in_Sentence        12
```

```
            Lip_Shape                                    8
        }
    }
}
```

Block 10 of FIG. 1 corresponds to the first 32 bits which specify a start of sentence code, and the following 10 bits that provide a sentence ID. The next bit indicates whether the sentence comprises a silence or voiced information, and if it is a silence, the next 12 bits specify the duration of the silence (block 11). Otherwise, the data that follows, as shown in block 13 provides information as to whether the Gender flag should be set in the synthesizer (1 bit), and whether the Age flag should be set in the synthesizer (1 bit). If the previously entered configuration parameters have set the Video_Enable flag to 0 and the Speech_Rate_Enable flag to 1 (block 14 of FIG. 1), then the next 4 bits indicate the speech rate. This is shown by block 14 of FIG. 1. Thereafter, the next 12 bits indicate the number of text bytes that will follow. This is shown by block 16 of FIG. 1. Based on this number, the subsequent stream of 8 bit bytes is read as the text input (per block 17 of FIG. 1) in the "for" loop that reads TTS_Text. Next, if the Video_Enable flag has been set by the previously entered configuration parameters (block 18 in FIG. 1), then the following 42 bits provide the silence duration (16 bits) the Position_in_Sentence (16 bits) and the Offset (10 bits), as shown in block 19 of FIG 1. Lastly, if the Lip_Shape_Enable flag has been set by the previously entered configuration parameters (block 20), then the following 51 bits provide information about lip shapes (block 21). This includes the number of lip shapes provided (10 bits), and the Lip_Shape_Time_in_Sentence (16 bits) if the Prosody_Enable and the Dur_Enable flags are set. If the Prosody_Enable flag is set but the Dur_Enable flag is not set, then the next 13 bits specify the Lip_shape_Phonem_Number_in_Sentence. If the Prosody_Enable flag is not set, then the next 12 bits provide the Lip_Shaper_letter_Number_in_Sentence information. The sentence ends with a number of lip shape specifications (8 bits each) corresponding to the value provided by Number_of_Lip_Shape field.

MPEG 4 provides for specifying phonemes in addition to specifying text. However, what is contemplated is to specify one pitch specification, and 3 energy specification, and this is not enough for high quality speech synthesis, even if the

synthesizer were to interpolate between pairs of pitch and energy specifications. This is particularly unsatisfactory when speech is aimed to be slow and rich is prosody, such as when singing, where a single phoneme may extend for a long time and be characterized with a varying prosody.

Summary of the Invention

An enhanced system is achieved which can specify that the stream of bits that follow corresponds to phonemes and a plurality of prosody information, including duration information, that is specified for times within the duration of the phonemes. Illustratively, such a stream comprises a flag to enable a duration flag, a flag to enable a pitch contour flag, a flag to enable an energy contour flag, a specification of the number of phonemes that follow, and, for each phoneme, one or more sets of specific prosody information that relates to the phoneme, such as a set of pitch values and their durations or temporal positions.

Brief Description of the Drawing

FIG. 1 visually represents signal components that may be applied to a speech synthesizer; and

FIG. 2 visually represents signal components that may be added, in accordance with the principles disclosed herein, to augment the signal represented in FIG. 1

Detailed Description

In accordance with the principles disclosed herein, instead of relying on the synthesizer to develop pitch and energy contours by interpolating between a supplied pitch and energy value for each phoneme, a signal is developed for synthesis which includes any number of prosody parameter target values. This can be any number, including 0. Moreover, in accordance with the principles disclosed herein, each prosody parameter target specification (such as amplitude of pitch or energy) is associated with a duration measure or time specifying when the target has to be reached. The duration may be absolute, or it may be in the form of offset from the beginning of the phoneme or some other timing marker.

A stream of data that is applied to a speech synthesizer in accordance with this invention may, illustratively, be one like described above, augmented with the following stream, inserted after the TTS_Text readings in the "for (j=0; j<Length_of_Text; j++)" loop. FIG. 2 provides a visual presentation of such a stream of bits that, correspondingly, is inserted following block 16 of FIG. 1.

```
if (Prosody_Enable) {
    Dur_Enable                                          1
    F0_Contour_Enable                                   1
    Energy_Contour_Enable                               1
    Number_of_Phonemes                                  10
    Phonemes_Symbols_length                             13
    for (j=0;j<Phoneme_Symbols_Length; j++)
        Phoneme_Symbols                                 8
    for (j=0; j<Number_of_Phonemes; j++) {
        if(Dur_Enable)
            Dur_each_Phoneme                            12
        if (F0_Contour_Enable) {
            num_F0                                      5
            for (j=0; ,<num_FO; j++) {
                F0_Countour_Each_Phoneme                 8
                F0_Countour_Each_Phoneme_time           12
                }
            }
        }
        if (Energy_Contour_Enable)
            Energy_Countour_Each_Phoneme                24
        }
    }
```

Proceeding to describe the above, if the Prosody_Enable flag has been set by the previously entered configuration parameters (block 30 in FIG. 2), the first bit in the bit stream following the reading of the text is a duration enable flag, Dur_Enable, which is 1

bit. This is shown by block 31. Following the Dur_Enable bit comes a one bit pitch enable flag, F0_Enable, and a one bit energy contour enable flag, Energy_Contour_Enable (blocks 32 and 33). Thereafter, 10 bits specify the number of phonemes that will be supplied (block 34) and the following 13 bits specify the number of 8 bit bytes that are required to be read (block 35) in order to obtain the entire set of phoneme symbols. Thence, for each of the specified phoneme symbols, a number of parameters are read as follows. If the Dur_Enable flag is set (block 37), the duration of the phoneme is specified in a 12 bit field (block 38). If the F0_Contour_Enable flag is set (block 39), then the following 5 bits specify the number of pitch specifications (block 40), and based on that number, pitch specifications are read in fields of 20 bits each (block 41). Each such field comprises 8 bits that specify the pitch, and the remaining 12 bits specify duration, or time offset. Lastly, if the Energy_Contour_Enable flag is set (block 42), the information about the energy contours is read in the manner described above in connection with the pitch information (block 43).

It should be understood that the collection and sequence of the information presented above and illustrated in FIG. 2 is merely that: illustrative. Other sequences would easily come to mind of a skilled artisan, and there is no reason why other information might not be included as well. For example, the sentence "hello world" might be specified by the following sequence:

Phoneme	Stress	Duration	Pitch and Energy Specs.
#	0	180	
h	0	50	P118@0 P118@24 A4096@0
e	3	80	
l	0	50	P105@19 P118@24
o	1	150	P117@91 P112@141 P137@146
#	1		
w	0	70	A4096@35
o			
R	1	210	P133@43 P84@54 A3277@105 A3277@210
l	0	50	P71@50 A3077@25 A2304@80

d	0	38+40	A4096@20 A2304@78
#			
*	0	20	P7@20 A0@20

It may be noted that in this sequence, each phoneme is followed by the specification for the phone, and that a stress symbols is included. A specification such as P133@43 in association with phoneme "R" means that a pitch value of 133 is specified to begin at 43 msec following the beginning of the "R" phoneme. The prefix "P" designates pitch, and the prefix "A" designates energy, or amplitude. The duration designation "38+40" refers to the duration of the initial silence (the closure part) of the phoneme "d," and the 40 refers to the duration of the release part that follows in the phoneme "d." This form of specification is employed in connection with a number of letters that consist of an initial silence followed by an explosive release part (e.g. the sounds corresponding to letters p, t, and k). The symbol "#" designates an end of a segment, and the symbol "*" designates a silence. It may be noted further that a silence can have prosody specifications because a silence is just another phoneme in a sequence of phonemes, and the prosody of an entire word/phrase/sentence is what is of interest. If specifying pitch and/or energy within a silence interval would improve the overall pitch and/or energy contour, there is no reason why such a specification should not be allowed.

It may be noted still further that allowing the pitch and energy specifications to be expressed in terms of offset from the beginning of the interval of the associated phoneme allows one to omit specifying any target parameter value at the beginning of the phoneme. In this manner, a synthesizer receiving the prosody parameter specifications will generate, at the beginning of a phoneme, whatever suits best in the effort to meet the specified targets for the previous and current phonemes.

An additional benefit of specifying the pitch contour as tuples of amplitude and time offset of duration is that a smaller amount of data has to be transmitted when compared to a scheme that specifies amplitudes at predefined time intervals.

We Claim:

1. A method for generating a signal rich in prosody information comprising:

a first step including in said signal a plurality of phoneme symbols,

a second step including in said signal a desired duration of each of said phoneme symbols,

a third step including at least one target prosody parameter value within a duration for at least one of said phonemes at a time offset from the beginning of the duration of said phoneme that is greater than zero and less than the duration of said phoneme.

2. The method of claim **1** where said prosody parameter is pitch.

3. The method of claim **1** where said prosody parameter is energy.

4. The method of claim **1** where said third step includes target values for both pitch and energy.

5. The method of claim **1** where at least some of the phonemes have no prosody parameter targets specified for the beginning of the durations of said at least some of the phonemes.

6. The method of claim **1** where timing of said prosody parameter target specifications are expressed in terms of durations.

7. The method of claim **1** where timing of said prosody parameter target specifications are expressed in terms of time offsets from the beginning of durations of phonemes.

8. The method of claim **1** where at least some silence intervals have one or more prosody parameter target specifications.

9. The method of claim **1** where the format of said signal is:

```
Dur_Enable
F0_Contour_Enable
Energy_Contour_Enable
Number_of_Phonemes
Phonemes_Symbols_length
for (j=0;j<Phoneme_Symbols_Length; j++)
    Phoneme_Symbols
for (j=0; j<Number_of_Phonemes; j++) {
    if(Dur_Enable)
        Dur_each_Phoneme
    if (F0_Contour_Enable) {
        num_F0
        for (j=0; ,<num_FO; j++) {
            F0_Countour_Each_Phoneme
            F0_Countour_Each_Phoneme_time
            }
        }
    }
    if (Energy_Contour_Enable)
        Energy_Countour_Each_Phoneme
    }
```

10. The method of claim **9** where said signal also includes text specifications.

11. The method of claim **1** where the format of said signal is:

```
Dur_Enable                                  1
F0_Contour_Enable                           1
Energy_Contour_Enable                       1
Number_of_Phonemes                          10
Phonemes_Symbols_length                     13
for (j=0;j<Phoneme_Symbols_Length; j++)
```

```
    Phoneme_Symbols                                          8
  for (j=0; j<Number_of_Phonemes; j++) {
    if(Dur_Enable)
      Dur_each_Phoneme                                       12
    if (F0_Contour_Enable) {
      num_F0                                                 5
      for (j=0; ,<num_FO; j++) {
        F0_Countour_Each_Phoneme                             8
        F0_Countour_Each_Phoneme_time                        12
        }
      }
    if (Energy_Contour_Enable)
      Energy_Countour_Each_Phoneme                           24
    }
```

where the numbers correspond to the number of bits.

12. The method of claim **1** the format of said signal is:

```
  TTS_Sentence_Start_Code                                    32
  TTS_Sentence_ID                                            10
  Silence                                                    1
  if (Silence)
    Silence_Duration                                         12
  else  {
    if (Gender_Enable)
      Gender                                                 1
    if (Age_Enable)
      Age                                                    3
    if (!Video_Enable & Speech_Rate_enable)
      Speech_Rate                                            4
    Length_of_Text                                           12
```

```
For (j=0; j<=Length_of_Text; j++)
    TTS_Text                                        8
if (Prosody_Enable) {
    Dur_Enable                                      1
    F0_Contour_Enable                               1
    Energy_Contour_Enable                           1
    Number_of_Phonemes                              10
    Phonemes_Symbols_length                         13
    for (j=0;j<Phoneme_Symbols_Length; j++)
        Phoneme_Symbols                             8
    for (j=0; j<Number_of_Phonemes; j++) {
        if(Dur_Enable)
            Dur_each_Phoneme                        12
        if (F0_Contour_Enable) {
            num_F0                                  5
            for (j=0; ,<num_FO; j++) {
                F0_Countour_Each_Phoneme            8
                F0_Countour_Each_Phoneme_time       12
               }
            }
        }
        if (Energy_Contour_Enable)
            Energy_Countour_Each_Phoneme            24
        }
        }
 if (Video_Enable) {
    if (Dur_Enable) {
        Sentence_Duration                           16
        Postion_in_Sentence                         16
        Offset                                      10
        }
```

```
            }
        if (Lip_Shape_Enable) {
            Number_of_Lip_Shape                                    10
            for (j=0; j<Number_of_Lip_Shape; j++) {
                If (Prosody_Enable) {
                    If (Dur_Enable)
                        Lip_Shape_Time_in_Sentence                 16
                    Else
                        Lip_Shape_Phoneme_Number_in_Sentence       13
                }
                else
                    Lip-Shape_Letter_Number_in_Sentence            12
                Lip_Shape                                          8
            }
        }
    }
```

where the numbers correspond to the number of bits.

Abstract

An enhanced system is achieved by allowing bookmarks which can specify that the stream of bits that follow corresponds to phonemes and a plurality of prosody information, including duration information, that is specified for times within the duration of the phonemes. Illustratively, such a stream comprises a flag to enable a duration flag, a flag to enable a pitch contour flag, a flag to enable an energy contour flag, a specification of the number of phonemes that follow, and, for each phoneme, one or more sets of specific prosody information that relates to the phoneme, such as a set of pitch values and their durations.

FIG. 1

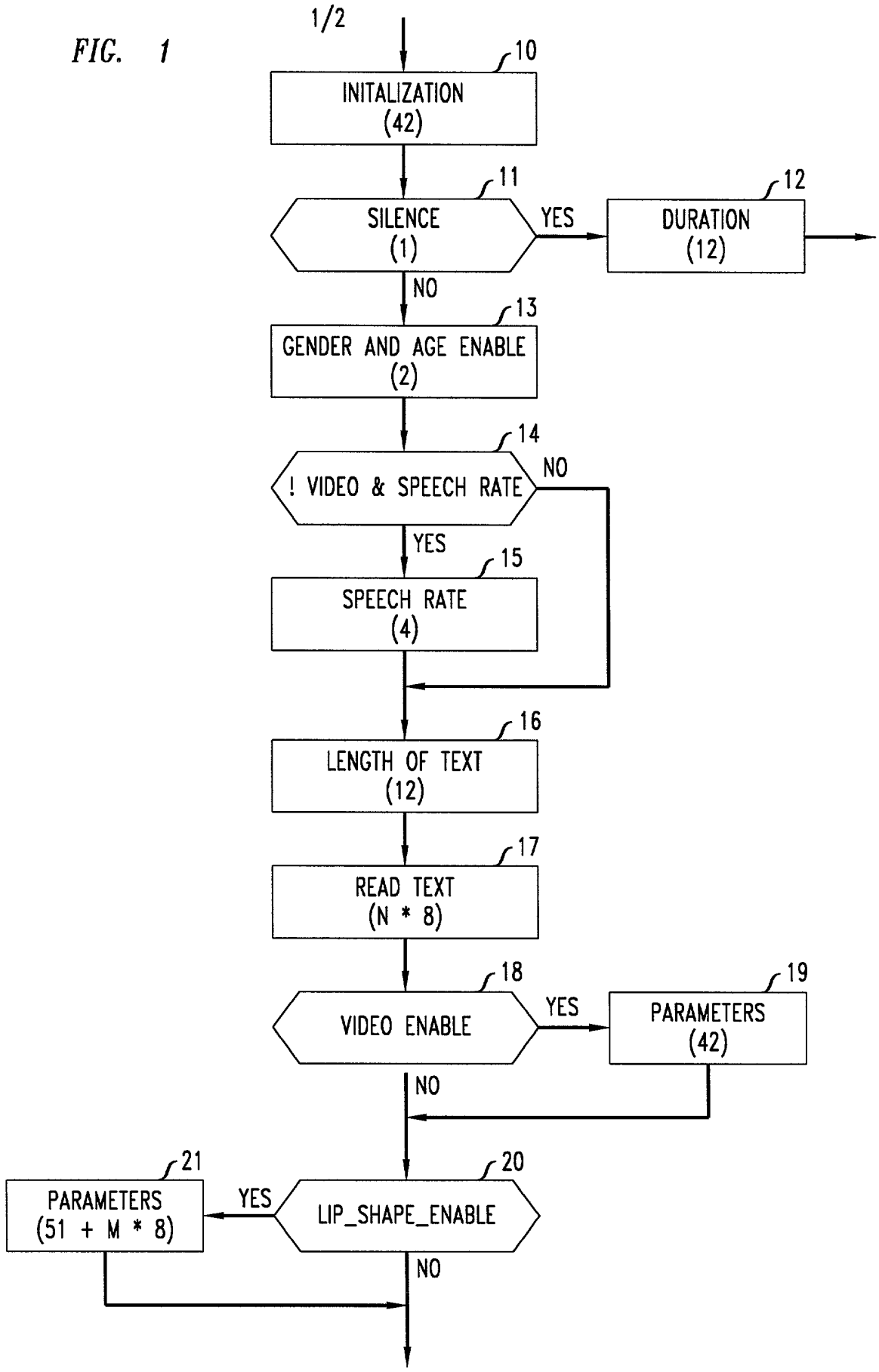
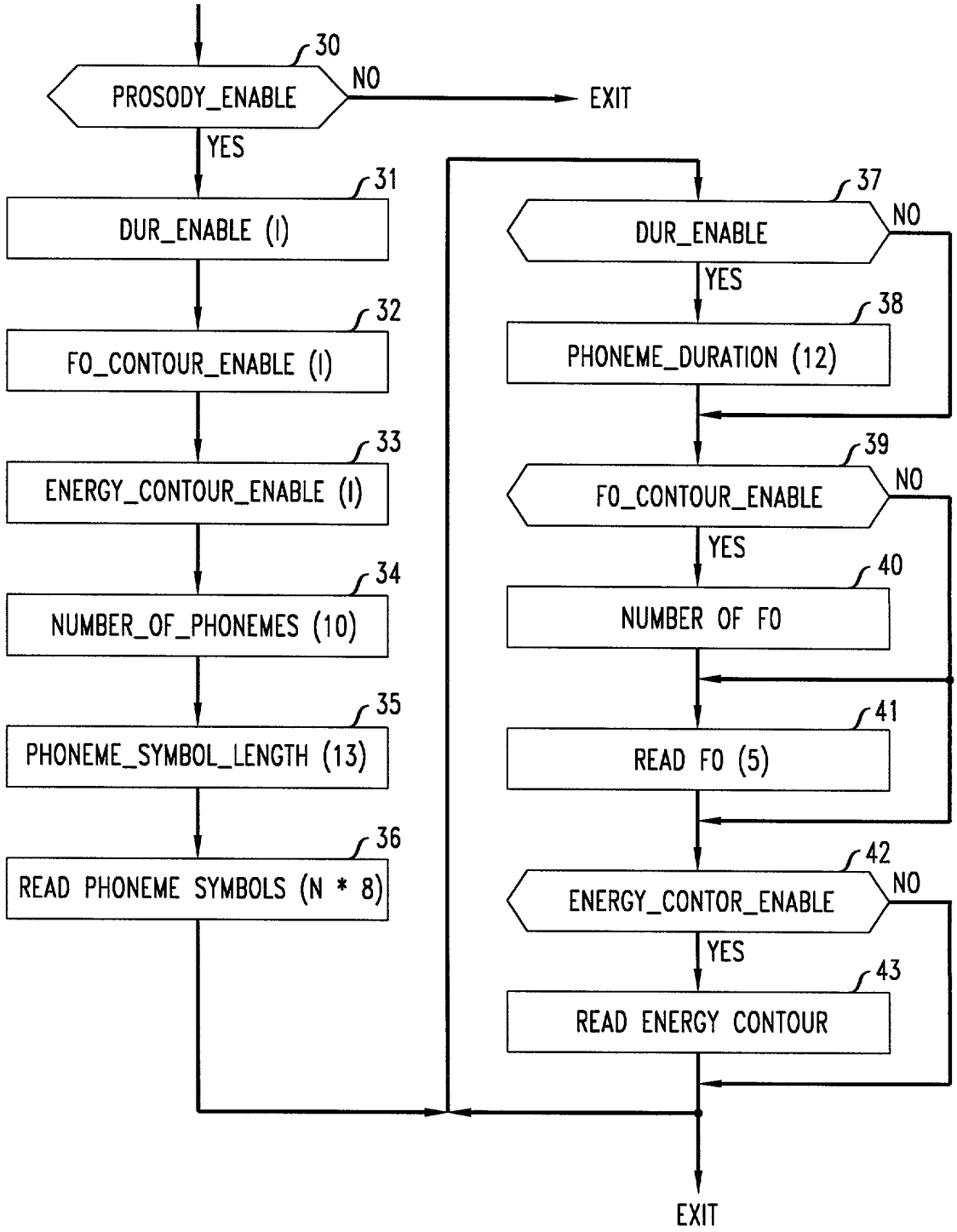


FIG. 2


30
PROSODY_ENABLE
NO
EXIT
YES
31
DUR_ENABLE (1)
32
F0_CONTOUR_ENABLE (1)
33
ENERGY_CONTOUR_ENABLE (1)
34
NUMBER_OF_PHONEMES (10)
35
PHONEME_SYMBOL_LENGTH (13)
36
READ PHONEME SYMBOLS (N * 8)
37
DUR_ENABLE
NO
YES
38
PHONEME_DURATION (12)
39
F0_CONTOUR_ENABLE
NO
YES
40
NUMBER OF F0
41
READ F0 (5)
42
ENERGY_CONTOR_ENABLE
NO
YES
43
READ ENERGY CONTOUR
EXIT

IN THE UNITED STATES
PATENT AND TRADEMARK OFFICE

Declaration and Power of Attorney

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled **Advanced TTS for Facial Animation** the specification of which was filed on January 27, 1999, as application Serial No. To be provided.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by an amendment, if any, specifically referred to in this oath or declaration.

I acknowledge the duty to disclose all information known to me which is material to patentability as defined in Title 37, Code of Federal Regulations, 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, 119 of any foreign application(s) for patent or inventors' certificate listed below and have also identified below any foreign application for patent or inventors' certificate having a filing date before that of the application on which priority is claimed:

None

I hereby claim the benefit under Title 35, United States Code, 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, 112, we acknowledge the duty to disclose all information known to us to be material to patentability as defined in Title 37, Code of Federal Regulations, 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

provisional application No. 60/073185, filed January 30, 1998, titled "Advanced TTS For Facial Animation," and
provisional application No. 60/082,393, filed April 20, 1998, titled "FAP Definition Syntax for TTS Input."

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

I hereby appoint the following attorney(s) with full power of substitution and revocation, to prosecute said application, to make alterations and amendments therein, to receive the patent, and to transact all business in the Patent and Trademark Office connected therewith:

Samuel H. Dworetsky	(Reg. No. 27873)
Thomas A. Restaino	(Reg. No. 33444)
Jose de la Rosa	(Reg. No. 34810)
Michele L. Conover	(Reg. No. 34962)
Robert B. Levy	(Reg. No. 28234)
Alfred G. Steinmetz	(Reg. No. 22971)
Benjamin S. Lee	(Reg. No. 42878)

I also appoint Henry T. Brendzel (Reg. No. 26,844) and William Ryan (Reg. No. 24,434) as associate attorneys, with full power to prosecute said application, to make alternations and amendments therein, and to transact all business in the Patent and Trademark Office connected therewith.

Please address all correspondence to Mr. S. H. Dworetsky, AT&T Corp., P.O. Box 4110, Middletown, New Jersey 07748. Telephone calls should be made to Henry T. Brendzel at (973) 467-2025.

Full name of joint inventor: Mark Charles Beutnagel

Inventor's signature _____ Date _____
Residence: Mendham, Morris County, NJ
Citizenship: USA
Post Office Address: 18 Mountain Avenue
Mendham, NJ 07945

Full name of joint inventor: Joern Osterman

Inventor's signature _____ Date _____
Residence: Red Bank, Monmouth County, NJ
Citizenship: Germany
Post Office Address: 72 Walnut Avenue
Red Bank, NJ 07701

Full name of joint inventor: Schuyler Reynier Quackenbush

Inventor's signature _____ Date _____
Residence: Westfield, Union County, NJ

Citizenship: USA
Post Office Address: 744 Tamaques Way
Westfield, NJ 07090